Bagging linear sparse Bayesian learning models for variable selection in cancer diagnosis

Devos, Andy; Van Huffel, Sabine; Lu, Chuan; Suykens, Johan A. K.; Arus, Carles

Published in:
IEEE Transactions on Information Technology in Biomedicine

DOI:
10.1109/TITB.2006.889702

Publication date:
2007

Citation for published version (APA):
Devos, A., Van Huffel, S., Lu, C., Suykens, J. A. K., & Arus, C. (2007). Bagging linear sparse Bayesian learning models for variable selection in cancer diagnosis. *IEEE Transactions on Information Technology in Biomedicine*, *11*(3), 338-347. https://doi.org/10.1109/TITB.2006.889702

Bagging linear sparse Bayesian learning models for variable selection in cancer diagnosis

Chuan Lu, Andy Devos, Johan A. K. Suykens, *Senior Member, IEEE,* Carles Arús, and Sabine Van Huffel, *Senior Member, IEEE*

***Abstract*— This work investigates variable selection and classification for biomedical datasets with a small sample size and a very high input dimension. The sequential sparse Bayesian learning methods with linear bases are used as the basic variable selection algorithm. Selected variables are fed to the kernel based probabilistic classifiers: Bayesian least squares support vector machines (LS-SVMs) and relevance vector machines (RVMs). We employ the bagging techniques for both variable selection and model building in order to improve the reliability of the selected variables and the predictive performance. This modelling strategy is applied to real-life medical classification problems, including two binary cancer diagnosis problems based on microarray data and a brain tumor multiclass classification problem using spectra acquired via magnetic resonance spectroscopy. The work is experimentally compared to other variable selection methods. It is shown that the use of bagging can improve the reliability and stability of both variable selection and model prediction.**

***Index Terms*— Variable selection, sparse Bayesian learning, bagging, kernel based probabilistic classifiers, microarray, magnetic resonance spectroscopy (MRS).**

I. Introduction

Recent advances in technologies such as microarrays and magnetic resonance (MR) have facilitated the collection of genomic, proteomic and metabolic data that can be used for medical decision support. For example, DNA microarrays enable us to simultaneously monitor the expression of thousands of genes [9][1]. It is then possible to compare the overall differences in gene expression between normal and diseased cells. Magnetic resonance spectroscopy (MRS) [16] is able to provide detailed chemical information about the metabolites presented in living tissue. In particular, *in vivo* proton MRS offers considerable potential for clinical applications, e.g. for brain tumor diagnosis [16][17].

Much attention has been paid to class prediction in the context of such new diagnostic tools, particularly for cancer diagnosis. The task is to classify and predict the category of a sample on the basis of its gene expression profile or the MRS spectrum. Conventional cancer diagnosis has been based on biopsy and examination by a pathologist of morphological appearance of stained tissue specimens in the microscope [9][1]. This method depends on the high expertise of the pathologists. Microarrays and MRS offer the hope that cancer classification can be objective and highly accurate, helping the clinicians to choose appropriate treatments. The challenge of classification using microarrays and MR spectra lies in: (1) the large number of input variables and a relatively small number of samples, and (2) the presence of noise and artefacts.

Kernel based methods are of particular interest for this task since they can deal with high dimensional data in nature and have been supported by both statistical learning theory and empirical results [10][29][7]. Despite the early success, the presence of a significant amount of irrelevant variables (features) or measurement noise might hamper the performance and interpretation of predictive models.

Variable selection (VS) is therefore used to identify the variables most relevant for classification. This is important for medical classification, as it can have an impact not only on the accuracy and complexity of the classifiers, but also on the economics of data acquisition. It is also helpful for understanding the underlying mechanisms of the disease. This could assist drug discovery and early diagnosis.

Several statistical and computational approaches to variable selection exist for classifying such data. The first approach is variable ranking followed by a selection (or filtering) step, usually accompanied with cross-validation (CV), to determine the number of variables to use in the classifier. The ranking criteria could be based on e.g. correlation, t-statistics and some multivariate methods such as recursive feature elimination (RFE) with support vector machines (SVMs) [10][7]. The second approach is the so-called wrapper approach, which searches for the optimal combination of variables according to some performance measures of the models [30][22][18].

The embedded approach combines the two tasks of variable selection and model fitting into one optimization procedure. The embedded SVM based algorithms typically reformulate the standard SVM optimization problem in order to select only a fixed number of variables. This can be done via imposing additional constraints and adopting objective functions such as generalization bound [29]. Nevertheless, these methods usually require an additional cross-validation step for choosing the predefined number of variables.

In [15] the Bayesian automatic relevance determination (ARD) algorithms, were exploited. This type of embedded methods can automatically determine the number of selected variables. However, they seem to be sensitive to a small

This work was supported by the projects of IUAP Phase V-22, of the KUL MEFISTO-666 and IDO/99/03, the FWO projects G.0407.02 and G.0269.02, and EU projects BIOPATTERN (FP6-2002-IST 508803), eTUMOUR (FP6-2002-LIFESCIHEALTH 503094) and HEALTHagents (FP6-2005-IST 027214). CL was supported by a doctoral grant of K.U.Leuven. AD was supported by an IWT grant (IWT-Vlaanderen).

CL is with Dept. of Computer Science, University of Wales, Aberystwyth, UK (e-mail: cul@aber.ac.uk). AD, JS and SVH are with SCD-SISTA, ESAT, Dept. of Electrical Engineering, Katholieke Universiteit Leuven, Belgium (e-mail: sabine.vanhuffel@esat.kuleuven.ac.be). CA is with Departament de Bioquímica i Biologia Molecular, Universitat Autonòma de Barcelona, 08193 Cerdanyola del Vallès, Spain.

permutation of the training set, rendering their results less reliable from a biological point of view.

In this paper, we explore an alternative method that follows the Bayesian ARD approach, and show how the reliability of the selected variables and classification performance can be improved by using bagging and feeding various bootstrap variables to various types of probabilistic models. Moreover, by utilizing the sparse Bayesian learning with logistic functions, this method requires no nuisance hyperparameters tuning.

The rest of the paper is organized as follows. Section II introduces sparse Bayesian learning algorithms. A brief review of the two kernel based probabilistic classification algorithms is given in Section III, namely Bayesian least squares support vector machines (BayLSSVM) and relevance vector machines (RVM). The bagging strategy for variable selection and modelling are proposed in Section IV. Section V lists the compared methods for variable selection and modelling. These methods are applied to the three cancer classification problems, as described in Section VI with results and the biological interpretation of some selected variables. Sections VII and VIII end the paper with a discussion and some conclusions.

II. Basic algorithm for variable selection

A. Sparse Bayesian learning

Supervised learning infers a functional relation $y \leftrightarrow f(\mathbf{x})$ from a training set $\mathbf{D} = \{\mathbf{x}_n, y_n\}_{n=1}^N$, with $\mathbf{x}_n \in \mathbb{R}^d$ and $y \in \mathbb{R}$. Sparse Bayesian learning (SBL) applies Bayesian ARD to models linear in their parameters so that sparse solutions (i.e. with many parameters equal to zero) can be obtained [25]. Its prediction on y given $\mathbf{x}$ can be based upon:

$$f(\mathbf{x};\mathbf{w}) = \sum_{m=0}^{M} w_m \phi_m(\mathbf{x}) = \mathbf{w}^{\mathbf{T}} \boldsymbol{\phi}(\mathbf{x}). \quad (1)$$

Two forms of basis functions $\phi_m(\mathbf{z})$ are considered here: $\phi_m(\mathbf{z}) = z_m, m = 1, \cdots, d$, (i.e. $\phi_m(\mathbf{z})$ as the original input variable), and $\phi_m(\mathbf{z}) = K(\mathbf{z}, \mathbf{x}_m), m = 1, \cdots, N$, where $K(.,.)$ denotes some symmetric kernel function. $\phi_0(\mathbf{z})$ is set to 1 in order to include an intercept (bias) term in the emerging model. The basic variable selection algorithm relies on the sparse Bayesian learning model using the first form of basis functions, termed the linear basis functions. In contrast, the relevance vector machines (RVMs) take the kernel representation for the basis function. The RVM will be revisited in Section III-B as a probabilistic classifier.

For a regression problem, the likelihood of the data for a sparse Bayesian learning model can be expressed as:

$$p(\mathbf{y}|\mathbf{w}, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp\{-\frac{1}{2\sigma^2}\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\|^2\}, \quad (2)$$

where σ^2 is the variance of the i.i.d. noise, the $N \times M$ design matrix $\boldsymbol{\Phi} = [\boldsymbol{\phi}(\mathbf{x}_1), \boldsymbol{\phi}(\mathbf{x}_2), \ldots, \boldsymbol{\phi}(\mathbf{x}_N)]^T$. The parameters $\mathbf{w}$ are given a Gaussian prior

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{m=0}^{M} \mathcal{N}(w_m|0, \alpha_m^{-1}) \quad (3)$$

where $\boldsymbol{\alpha} = \{\alpha_m\}$ is a vector of hyperparameters, with one hyperparameter α_m assigned to each model parameter w_m. As illustrated in [25][3], this is equivalent to using a regularization with a penalty of $\sum_m \log|w_m|$, which encourages sparsity. Given $\boldsymbol{\alpha}$, the posterior parameter distribution can be derived via the Bayes' rule

$$p(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha}, \sigma^2) = p(\mathbf{y}|\mathbf{w}, \sigma^2) p(\mathbf{w}|\boldsymbol{\alpha}) / p(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2), \quad (4)$$

which is also Gaussian, with variance and mean of

$$\boldsymbol{\Sigma} = (\sigma^2 \boldsymbol{\Phi}^T \boldsymbol{\Phi} + \mathbf{A})^{-1} \text{ and } \boldsymbol{\mu} = \sigma^{-2} \boldsymbol{\Sigma} \boldsymbol{\Phi}^T \mathbf{y}. \quad (5)$$

The hyperparameters $\boldsymbol{\alpha}$ can be estimated using type II maximum likelihood, in which the marginal likelihood is maximized. And the marginal likelihood can be computed by:

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2) &= \int_{-\infty}^{\infty} p(\mathbf{y}|\mathbf{w}, \sigma^2) p(\mathbf{w}|\boldsymbol{\alpha}) \mathbf{dw} \\ &= (2\pi)^{-N/2} |\mathbf{C}|^{-1/2} \exp(-\frac{1}{2}\mathbf{y}^T \mathbf{C}^{-1} \mathbf{y}) \quad (6) \end{aligned}$$

where $\mathbf{C} = \mathbf{B^{-1}} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^T$, with $\mathbf{B} = \sigma^{-2}\mathbf{I}$ and $\mathbf{A} = \mathrm{diag}(\alpha_1, \ldots, \alpha_M)$.

For binary classification problems, one can utilize the logistic function $g(a) = 1/(1 + e^{-a})$ [25]. The computation of the likelihood is based on the Bernoulli distribution:

$$p(\mathbf{y}|\mathbf{w}) = \prod_{n=1}^{N} g(f(\mathbf{x}_n;\mathbf{w}))^{y_n} [1 - g(f(\mathbf{x}_n;\mathbf{w}))]^{1-y_n}, \quad (7)$$

where $y_n \in \{0, 1\}$. There is no noise variance in this case, and a local Gaussian approximation is used to compute the posterior distribution of the weights $p(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha})$ and the marginal likelihood $p(\mathbf{y}|\boldsymbol{\alpha})$. For a given $\boldsymbol{\alpha}$, we can estimate the mean and variance of the weights ($\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$) by an iteratively reweighted least squares algorithm (e.g. the Newton-Raphson method [2]). The following expressions are exploited [25]:

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}} &= (\boldsymbol{\Phi}^T \mathbf{B} \boldsymbol{\Phi} + \mathbf{A})^{-1}, \\ \widehat{\boldsymbol{\mu}} &= \widehat{\boldsymbol{\Sigma}} \boldsymbol{\Phi}^T \mathbf{B} \widehat{\mathbf{y}}, \\ \text{and } \widehat{\mathbf{y}} &= \boldsymbol{\Phi}\widehat{\boldsymbol{\mu}} + \mathbf{B}^{-1}(\mathbf{y} - g(\boldsymbol{\Phi}\widehat{\boldsymbol{\mu}})), \end{aligned}$$

where $\mathbf{B} = \mathrm{diag}(\beta_1, \ldots, \beta_N)$, with $\beta_n = g(f(\mathbf{x}_n;\widehat{\boldsymbol{\mu}}))[1 - g(f(\mathbf{x}_n;\widehat{\boldsymbol{\mu}}))]$.

The optimization process, i.e. maximization of the marginal likelihood with respect to $\boldsymbol{\alpha}$ and possibly σ^2, can be performed efficiently using an iterative re-estimation procedure [25][3]. A fast sequential learning algorithm has also been introduced in [26], which enables us to efficiently process data of high dimensionality. We have adapted this algorithm to our applications, which will be detailed in the next subsection.

The most relevant variables for the classifier can be obtained from the resulting sparse solutions, if the original variables are taken as basis functions in the SBL model. This type of model is referred to as the linear SBL models in this paper.

B. Sequential sparse Bayesian learning algorithm

The sequential SBL algorithm [26] starts from a zero basis, adds or deletes a basis function at each iteration step, or updates a hyperparameter α_m until convergence.

For optimization of the hyperparameters $\boldsymbol{\alpha}$, the objective function uses the logarithm of the marginal likelihood $\mathcal{L}(\boldsymbol{\alpha}) = \log p(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2)$. It is shown in [26][3] that we may analyze the properties of $\mathcal{L}(\boldsymbol{\alpha})$ by decomposing it into the marginal

likelihood $\mathcal{L}(\boldsymbol{\alpha}_{-i})$ with $\boldsymbol{\phi}_i$ (the ith column of $\boldsymbol{\Phi}$) excluded, and the marginal likelihood $\ell(\alpha_i)$ including only $\boldsymbol{\phi}_i$. That is $\mathcal{L}(\boldsymbol{\alpha}) = \mathcal{L}(\boldsymbol{\alpha}_{-i}) + \ell(\alpha_i)$, where $\ell(\alpha_i) = \frac{1}{2}[\log\alpha_i - \log(\alpha_i + s_i) + \frac{q_i^2}{\alpha_i + s_i}]$, $s_i = \boldsymbol{\phi}_i^T \mathbf{C}_{-i}^{-1} \boldsymbol{\phi}_i$ and $q_i = \boldsymbol{\phi}_i^T \mathbf{C}_{-i}^{-1} \mathbf{y}$, $\mathbf{C}_{-i}$ is $\mathbf{C}$ with the contribution of $\boldsymbol{\phi}_i$ removed. Since s_i and q_i are independent of α_i, one can obtain a unique maximum of $\mathcal{L}(\boldsymbol{\alpha})$ with respect to α_i by setting the first derivative of $\ell(\alpha_i)$ to zero. The optimal values for α_i are:

$$\tilde{\alpha}_i = \frac{s_i^2}{q_i^2 - s_i}, \text{ if } q_i^2 > s_i, \quad (8)$$

$$\tilde{\alpha}_i = \infty, \text{ if } q_i^2 \leq s_i. \quad (9)$$

One convenient way to derive s_i and q_i is to utilize these expressions: $s_i = \frac{\alpha_i S_i}{\alpha_i - S_i}$, $q_i = \frac{\alpha_i Q_i}{\alpha_i - S_i}$, and when $\alpha_i = \infty$, $s_i = S_i$ and $q_i = Q_i$. In practice the two quantities S_i and Q_i are computed using the following equations:

$$S_i = \boldsymbol{\phi}_i^T \mathbf{C}^{-1} \boldsymbol{\phi}_i = \boldsymbol{\phi}_i^T \mathbf{B} \boldsymbol{\phi}_i - \boldsymbol{\phi}_i^T \mathbf{B} \boldsymbol{\Phi} \hat{\boldsymbol{\Sigma}} \boldsymbol{\Phi}^T \mathbf{B} \boldsymbol{\phi}_i \quad (10)$$

$$Q_i = \boldsymbol{\phi}_i^T \mathbf{C}^{-1} \hat{\mathbf{y}} = \boldsymbol{\phi}_i^T \mathbf{B} \hat{\mathbf{y}} - \boldsymbol{\phi}_i^T \mathbf{B} \boldsymbol{\Phi} \hat{\boldsymbol{\mu}}, \quad (11)$$

where $\boldsymbol{\Phi}$, $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\mu}}$ contain only the parts corresponding to the basis functions included in the model (with $\alpha_m < \infty$).

The marginal likelihood maximization algorithm jointly optimizes the weights and the hyperparameters $\{\alpha_m\}_{m=0}^{M_{\text{all}}}$, with M_{all} the maximum index for the basis functions. In case of linear basis functions, $M_{\text{all}} = d$; in case of kernel basis functions $M_{\text{all}} = N$. Define the complete set of possible indices for the basis functions as $\mathcal{I}_{\text{all}}$, containing the integer numbers from 0 to M_{all}. Our modified algorithm of SBL for classification utilizing the logistic function is as follows.

1) Initialize the model with only an intercept: $\alpha_0 < \infty$ (e.g. $\alpha_0 = (\mathbf{y}^T\mathbf{y}/N)^{-2}$), and $\forall m > 0, \alpha_m = \infty$. Initialize the index set of the bases in the model $\mathcal{I}_{\text{sel}} \leftarrow \{0\}$.
2) Given current $\boldsymbol{\alpha}$, estimate $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\mu}}$ using the IRLS algorithm for the logit model. Note that $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\mu}}$ are only related to the basis functions included in the current model, initially with only one scalar element. And $\boldsymbol{\Phi}$ starts with only one column vector $\boldsymbol{\phi}_0 = [1, \cdots, 1]_{1\times N}^T$.
3) Randomly select M_{out} bases with the indices of $\mathcal{I}_{\text{out}} \subseteq (\mathcal{I}_{\text{all}} \setminus \mathcal{I}_{\text{sel}})$. Set $\mathcal{I}_{\text{can}} \leftarrow \mathcal{I}_{\text{out}} \cup \mathcal{I}_{\text{sel}}$.
4) For each basis vector in the candidate sets $\boldsymbol{\phi}_m, m \in \mathcal{I}_{\text{can}}$, compute the value of s_m and q_m, find the optimal action with respect to each α_m, then calculate ∇_m, the corresponding change in marginal likelihood $\mathcal{L}(\boldsymbol{\alpha})$ after taking that action. The following rules are used:
   - If $q_m^2 > s_m$ and $\alpha_m < \infty$ (i.e. $\boldsymbol{\phi}_i$ is in the model), estimate $\tilde{\alpha}_m$ using (8), $\nabla_m = \frac{Q_m^2}{S_m + [\tilde{\alpha}_m^{-1} - \alpha_m^{-1}]^{-1}} - \log\{1 + S_m[\tilde{\alpha}_m^{-1} - \alpha_m^{-1}]\}$.
   - If $q_m^2 > s_m$ and $\alpha_m = \infty$, add $\boldsymbol{\phi}_m$ to the model, compute $\tilde{\alpha}_m$ using (8), $\nabla_m = \frac{Q_m^2 - S_m}{S_m} + \log\frac{S_m}{Q_m^2}$.
   - If $q_m^2 \leq s_m$ and $\alpha_m < \infty$, then delete $\boldsymbol{\phi}_m$, set $\tilde{\alpha}_m = \infty$, $\nabla_m = \frac{Q_m^2}{S_m - \alpha_m} - \log(1 - \frac{S_m}{\alpha_m})$.
5) Select one basis $m^* = \arg\max \nabla_m$, take the corresponding action, i.e. $\alpha_{m^*} \leftarrow \tilde{\alpha}_{m^*}$ and update $\boldsymbol{\Phi}$ and $\mathcal{I}_{\text{sel}}$.
6) If convergence is reached then stop, otherwise goto step 2).

The number of bases to be screened for updating α_m is the number of bases in the model plus M_{out}, the predefined number of randomly selected bases from those not used by the model. Although M_{out} should be chosen empirically over computational efficiency, quite a wide range of values may give satisfactory results. In our experiments, it is fixed to 100. Here the optimization procedure is considered to be converged when the maximum value of $|\log(\tilde{\alpha}_m/\alpha_m)|_{m\in\mathcal{I}_{\text{can}}}$ in step 4 is very small, e.g. lower than 10^{-6}.

However, we should also be aware of the uncertainty involved in the basis function selection, which might result from the existence of multiple solutions and the sensitivity of the algorithm to small perturbations of experimental conditions. Attempts to tackle this problem are for example bagging and committee machines. Here we will focus on the very simple bagging approach, which will be described in Section IV.

III. Kernel based probabilistic classifiers

Support Vector Machines (SVM) are now a state-of-the-art technique for pattern recognition [27]. A standard SVM classifier takes the form $y(\mathbf{x}) = \text{sign}[\mathbf{w}_{\text{f}}^T \boldsymbol{\varphi}(\mathbf{x}) + b]$ in the feature (primal) space with $\boldsymbol{\varphi}(.) : \mathbb{R}^d \rightarrow \mathbb{R}^{d_{\text{f}}}$, where d_{f} is the dimension of the feature space. It is inferred from data with binary targets $y_i \in \{\pm 1\}$ by solving the following optimization problem:

$$\min_{\mathbf{w}_{\text{f}}, b, \boldsymbol{\xi}} \mathcal{J}(\mathbf{w}_{\text{f}}, b, \boldsymbol{\xi}) = \frac{1}{2}\mathbf{w}_{\text{f}}^T \mathbf{w}_{\text{f}} + C\sum_{i=1}^{N} \xi_i, \quad (12)$$

subject to $y_i(\mathbf{w}_{\text{f}} \boldsymbol{\varphi}(\mathbf{x}_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \ldots, N$.

This can be conveniently solved in its dual formulation. It turns out that $f(\mathbf{x}; \mathbf{w}_{\text{f}}, b) = \mathbf{w}_{\text{f}}^T \boldsymbol{\varphi}(\mathbf{x}) + b = \sum_{i=1}^{N} a_i\, y_i\, K(\mathbf{x}, \mathbf{x}_i) + b$, where a_i is called a support value, and $K(\cdot, \cdot)$ is a chosen positive definite kernel. The most common kernels include linear kernels and radial basis function (RBF) kernels. Here we only considered models with linear kernels, defined as $K(\mathbf{x}, \mathbf{z}) = \mathbf{x}^T\mathbf{z}$.

A. Bayesian LS-SVM classifier

The LS-SVM is a least squares version of SVM, and is closely related to Gaussian processes and kernel Fisher discriminant analysis [23][24]. The training procedure for LS-SVM is reformulated as

$$\min_{\mathbf{w}_{\text{f}}, b, \mathbf{e}} \mathcal{J}(\mathbf{w}_{\text{f}}, b, \mathbf{e}) = \frac{1}{2}\mathbf{w}_{\text{f}}^T \mathbf{w}_{\text{f}} + \frac{\lambda}{2}\sum_{i=1}^{N} e_i^2, \quad (13)$$

subject to $y_i[\mathbf{w}_{\text{f}}^T \boldsymbol{\varphi}(\mathbf{x}_i) + b] = 1 - e_i, i = 1, \ldots, N$.

This optimization problem can be transformed and solved through a linear system in the dual space instead of a quadratic programming problem as for the standard SVM case [24]:

$$\left[\begin{array}{c|c} 0 & \mathbf{y}^T \\ \hline \mathbf{y} & \boldsymbol{\Omega} + \lambda^{-1}\mathbf{I} \end{array}\right] \left[\begin{array}{c} b \\ \hline \mathbf{a} \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline \mathbf{1} \end{array}\right] \quad (14)$$

where $\mathbf{a} = [a_1, \cdots, a_N]^T$, and $\mathbf{1} = [1 \cdots 1]^T$. The matrix $\boldsymbol{\Omega}$ is defined as $\boldsymbol{\Omega}_{ij} = y_i y_j \boldsymbol{\varphi}(\mathbf{x}_i)^T \boldsymbol{\varphi}(\mathbf{x}_j) = y_i y_j\, K(\mathbf{x}_i, \mathbf{x}_j)$.

In Bayesian LS-SVM (BayLSSVM) [24][28], the LS-SVM is integrated with the evidence framework [2], within which the regularization parameters λ is optimized by maximizing the posterior probability of the model. And the posterior class probabilities can be calculated incorporating the prior class probabilities via the Bayes' rule.

B. Relevance vector machines for classification

The RVM is a special case of SBL models, in which the basis functions are given by kernel functions of the same type as for SVMs. The sequential learning algorithm introduced in Section II-B is again applied to the optimization of RVMs. The predicted probability of being positive for a given input $\mathbf{x}_*$ can be computed using the logistic function:

$$p(y_* = 1|\mathbf{x}_*, \mathbf{y}, \boldsymbol{\alpha}) = \frac{1}{1 + e^{-\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_*)}} \tag{15}$$

No simulation results are reported for the models with nonlinear kernels in this paper. As linear classifiers perform sufficiently well for our problems, and nonlinear models have shown no improvement over the simple linear classifiers, according to the results of our preliminary experiments and some other studies on the same datasets [10][6].

IV. Bagging strategy

A. Bagging the selected variables and models

Bagging is a bootstrap ensemble method that generates individuals for its ensemble by training each classifier on a random redistribution of the training set [4]. Each classifier's training set is generated by randomly drawing, with replacement, the same number of examples as in the original training set.

It is shown that the bootstrap mean is approximately a posterior average of a quantity of interest [13]. Suppose a model is fitted to our training set $\mathbf{D}$, obtaining the prediction $f(\mathbf{x})$ at input $\mathbf{x}$. This prediction could be the latent outcome of e.g. a standard SVM model, or the predicted class probability of a probabilistic model. Bootstrap aggregation or bagging averages this prediction over a collection of bootstrap samples, thereby reducing its variance. For each bootstrap sample $\mathbf{D}^{*b}, b = 1, 2, \cdots, B$, we fit the model, giving prediction $f^{*b}(\mathbf{x})$. The bagging estimate is defined by

$$f_{\text{bag}}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} f^{*b}(\mathbf{x}). \tag{16}$$

The final class label will be decided by thresholding the bootstrap estimate of the class probability or the latent outcome. Bagging can push a good but unstable procedure a significant step towards optimality, which has been witnessed both empirically and theoretically [4][13].

An alternative bagging strategy is to bag only the predicted class labels and the final prediction will be given by voting. However, a reliable estimate of the class probability is essential for medical diagnosis. The prediction averaging strategy tends to produce bagged estimates with lower variance, especially for small B. Therefore, the prediction averaging strategy is preferred and advocated here.

The bagging strategy for variable selection and modelling is outlined below. Given a training set, B bootstrap data are randomly generated with replacement. For each bootstrap training set, one subset of variables is selected via the linear SBL logit models, followed by feeding these variables to a model of interest such as Bayesian LS-SVM. Then, B subsets of variables are chosen and B corresponding models built based on the B bootstrap training data. Given input $\mathbf{x}$, the class probability or the latent outcome for the bagged binary model will be the average of the B class probabilities or latent outcomes. The pseudo code for this is given in Algorithm 1. B is empirically set to 30 in our experiments.

B. Strategy for the multiclass classification problems

For multiclass classification problems, we reduce the k-class classification problem into $k(k-1)/2$ pairwise binary classification problems, which yield the conditional pairwise probability estimates for a given input $\mathbf{x}$. The conditional probabilities are then coupled to obtain the joint posterior probability for each class by using Hastie's method [12]. The final prediction of the class will be the one for which the highest joint probability is achieved. Accordingly the variables used should be the union of the $k(k-1)/2$ sets of variables that are used in the same number of binary classifiers. Bagging is applied to each binary classification individually. Only the mean predicted probabilities from the bagged binary classifiers are coupled in order to get the final joint posterior probability for the multiclass classification problems.

V. Compared methods

In order to see the potential performance gain of using our proposed methods, we have also assessed the performance of some reference methods. We denote the proposed variable selection approach as "LinSBL+Bag" (see Algorithm 1), which bags the variables selected from the linear SBL logit models. Accordingly the model fitting and prediction will be "bagged" as well. Its counterpart method "LinSBL" forms a classifier using only one subset of variables selected from a single linear SBL logit model, which is based on the whole training set without bootstrap repetition.

Other two variable selection methods adopted for comparison involve variable ranking followed by selecting N_{v} variables with the highest ranks. One is the popular SVM based recursive feature elimination (RFE) method [10]. The idea of this method is to eliminate recursively the variable which contributes the least in the SVM model, and then rank the variables based on the reverse order of their elimination. The contribution of the mth variable is evaluated by means of the change in the cost function $\nabla \mathcal{J}_m$ caused by removing the mth variable. When a linear kernel is used, $\nabla \mathcal{J}_m = w_m^2$ with w_m the corresponding weight in the linear SVM model: $w_m = \Sigma_{i=1}^{N} a_i x_{im} y_i$.

The variables can also be ranked using Fisher's criterion [2], which is a measure of the correlation between the variables and the class labels. For a binary classification, the Fisher discriminant criterion for an individual variable is given by $(\mu_{m,+} - \mu_{m,-})^2/(\sigma_{m,-}^2 + \sigma_{m,+}^2)$, where $\mu_{m,+}$ and $\mu_{m,-}$ are the means of variable m within the positive and negative class, respectively, and $\sigma_{m,+}$ and $\sigma_{m,-}$ are the standard deviations of the variable within each class. The larger the Fisher's criterion, the higher the ranking of the variable.

N_{v} is tuned by 10-fold cross-validation using SVMs. A coarse-to-fine strategy is utilized to search for N_{v} within a range of possible values. The N_{v} with the lowest 10-fold CV error rate were selected, and the tie breaking rule is to choose

Input : Training set $\boldsymbol{D}$, number of bootstrap samples B, classifier L (such as SVM, LS-SVM, RVM).
Output : model ensemble $f_{\rm bag}$.

for $b = 1$ **to** B **do**
  $\boldsymbol{D}^{*b}$=bootstrap sample from $\boldsymbol{D}$;
  $\boldsymbol{V}^{*b}$= **LinSBL**$(\boldsymbol{D}^{*b})$ /*Select subset of variables $\boldsymbol{V}^{*b}$ via linear sparse Bayesian learning */;
  $f^{*b} = L(\boldsymbol{D}^{*b}, \boldsymbol{V}^{*b})$;
end
$f_{\rm bag}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} f^{*b}(\mathbf{x})$;

Algorithm 1: Ensemble modelling using LinSBL+Bag for variable selection.

Input : Training set $\boldsymbol{D}$, variable ranking method **Ranking** (such as RFE and Fisher), classifier L (such as SVM, LS-SVM, RVM).
Output: model f.

$\boldsymbol{R} =$ **Ranking**$(\boldsymbol{D})$ /*Get a list of variable rankings*/;
$\boldsymbol{V} = \mathrm{SVMCV}(\boldsymbol{D}, \boldsymbol{R})$ /*Select $\boldsymbol{V}$ which consists of $N_{\rm v}$ variables with the highest ranking by SVM CV */;
$f = L(\boldsymbol{D}, \boldsymbol{V})$;

Algorithm 2: Modeling using Ranking+CV for variable selection.

the smallest number of variables. These two VS methods are denoted by "RFE+CV" and "Fisher+CV", respectively.

Note that no bagging has been applied for these reference methods. Our preliminary experiments show that the effect in bagging the models is more prominent when the variables vary among different bootstrap data. However, it becomes too time consuming to bootstrap variable selection with the methods of "Ranking + CV" (see Algorithm 2).

Concerning the modelling techniques, in addition to the advocated probabilistic models, we use the standard linear SVM classifier as a baseline model. We fix the regularization hyperparameter of SVM to 10^6, high enough to keep the training error low. Unlike other probabilistic models, the final SVM classifiers do not generate naturally the probability output. Hence, for the multiclass classification problem, the final predicted class labels are decided by voting, using the pairwise binary SVM classification results. We also compare the kernel-based models with the decision tree models obtained from C4.5 [19], which is a classical machine learning algorithm. The output of the bagged C4.5 models is given by voting.

VI. Experiments

A. Experimental settings

The generalization performance of the models in conjunction with variable selection was evaluated by 30 runs of randomized holdout cross-validation. For each run, a fixed proportion of data were taken for training and the rest for test, and the spiltting of the dataset was random stratified.

We applied a full cross-validation, where the variable selection was conducted prior to each model fitting process for each realization of the training data. An incomplete cross-validation, i.e. a cross-validation after variable selection may lead to a serious underestimation of the prediction error [21].

The performance is measured by the mean accuracy (Acc) and the mean area under the ROC curve (AUC) [11] and the corresponding standard error (SE) of the mean.

Our matlab programs used for these experiments were built upon several toolboxes, including the SparseBayes V1.0 [1] (with modifications) for the sequential sparse Bayesian learning, the Spider[2] for RFE, SVM and C4.5 modelling, and LS-SVMlab [3] for Bayesian LS-SVM modelling.

Note that in our experiments, all classifiers were tested with the same series of variable selection techniques.

B. Binary cancer classification based on microarray data

Two benchmark binary cancer classification problems based on DNA microarray data have been considered. The first problem aims to discriminate between two types of leukemia: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The dataset includes 72 samples (47 of AML and 25 of ALL) with 7129 gene expression values [4] [9][10].

The second problem aims to differentiate tumors from normal tissues using colon cancer data. The dataset contains information of 40 tumor and 22 normal colon tissues. Each tissue is associated with 2000 gene expression values [5] [1][10].

All microarray data have been normalized to zero mean and unit variance. Each realization of the training set contains 50 data points, the test set includes the rest of the data of size 22 and 12 in leukemia and colon cancer data, respectively. These two problem cases are both linearly separable, however, they both have very high dimensionality and small sample size.

By default, the class priors were set to the proportions of the classes in the training set in the binary classifications. The accuracy and the AUC for the leukemia and colon cancer classification problems are reported in Table I and Table II, respectively, where the highest value of accuracy or AUC for each type of classifier (in row) is indicated in bold. The mean number of selected variables ($\overline{N_{\rm v}}$) for each variable selection (VS) method within one trial is also given. Note that, only the accuracy measure is reported for C4.5 decision tree classifiers, as the latent outcome is not available for C4.5 models for the ROC analysis.

We see that RFE+CV and LinSBL selected less variables and resulted in a consistently lower test performance than Fisher+CV and LinSBL+Bag. Models using only a small subset of variables selected by Fisher+CV and LinSBL+Bag achieved a similar test result as the models without variable selection. Comparing different C4.5 models in terms of accuracy, the bagged C4.5 performed significantly better than the other single C4.5 models. All the other kernel-based models performed better than these decision trees. Among the kernel based models, the use of variable selection is considered more significant than the choice of any particular type of classifier with regard to the model performance.

Additionally, the test performance from the LinSBL and LinSBL+Bag were compared via paired t-tests for each classifier. The p-values of comparison on AUC are all $< 10^{-4}$ for the leukemia data, and all < 0.015 for the colon data.

[1] http://research.microsoft.com/mlp/RVM/SparseBayesV1.00.tar.gz
[2] http://www.kyb.tuebingen.mpg.de/bs/people/spider/index.html
[3] http://www.esat.kuleuven.ac.be/sista/lssvmlab/
[4] available online at www.genome.wi.mit.edu/MPR/data_set_ALL_AML.html
[5] available online at microarray.princeton.edu/oncology/affydata/index.html

TABLE I

TEST RESULTS FOR LEUKEMIA CANCER CLASSIFICATION.

VS method	All		Fisher+CV		RFE+CV		LinSBL		LinSBL+Bag	
$\overline{N}_{\rm v}$	7129		43.30		5.10		3.50		49.10	
Classifier	Acc(%)	AUC(%)	Acc(%)	AUC(%)	Acc(%)	AUC(%)	Acc(%)	AUC	Acc(%)	AUC(%)
SVM	**95.61**	**99.29**	92.42	97.29	90.30	94.88	90.45	94.43	92.42	98.01
	±0.73	±0.23	±0.89	±0.64	±1.19	±1.17	±0.92	±0.94	±0.92	±0.46
BayLSSVM	89.85	**98.78**	92.73	97.62	90.00	95.01	88.94	93.50	**93.79**	98.48
	±1.05	±0.34	±1.04	±0.59	±1.29	±1.11	±1.02	±1.40	±1.02	±0.41
RVM	90.15	96.22	93.03	97.26	90.45	94.93	89.85	93.72	**93.18**	**98.24**
	±1.34	±0.77	±0.85	±0.61	±1.24	±1.10	±1.17	±1.17	±1.00	±0.47
C4.5	85.00		86.82		86.82		88.79		**92.58**	
	±1.00		±1.04		±1.12		±0.85		±0.84	

TABLE II

TEST RESULTS FOR COLON CANCER CLASSIFICATION.

VS method	All		Fisher+CV		RFE+CV		LinSBL		LinSBL+Bag	
$\overline{N}_{\rm v}$	2000		114.30		9.73		6.50		107.17	
Classifier	Acc(%)	AUC(%)	Acc(%)	AUC(%)	Acc(%)	AUC(%)	Acc(%)	AUC(%)	Acc(%)	AUC(%)
SVM	81.94	85.31	80.28	85.83	81.36	83.54	76.39	81.67	**86.11**	**87.71**
	±1.80	±2.22	±2.10	±2.26	±2.18	±3.10	±2.03	±2.23	±1.73	±1.93
BayLSSVM	85.00	88.65	**85.28**	**89.27**	81.11	86.15	76.67	84.48	84.44	89.06
	±1.73	±1.71	±1.56	±1.60	±2.11	±2.79	±2.08	±2.09	±1.65	±1.78
RVM	83.61	87.71	**83.61**	86.98	81.67	86.46	73.06	82.40	79.17	**87.71**
	±1.82	±2.16	±1.73	±2.17	±2.30	±2.75	±2.41	±2.23	±1.99	±1.88
C4.5	73.61		76.39		76.67		73.33		**84.17**	
	±2.61		±2.08		±2.62		±1.82		±1.60	

C. Classification of brain tumors based on MRS data

The method has also been applied to a multiclass classification problem of brain tumors using short echo time ^{1}H MRS data. The dataset consists of 205 spectra in the frequency domain. The full spectrum (a row vector of magnitude values) has been normalized to unit norm. Only the frequency region of interest from 4.17 to 0 ppm (a measure of the chemical shift in a field independent frequency scale) was used in this study, corresponding to 138 input variables. The dataset contains the records from four major types of brain tumor: meningiomas (Class 1, 57 spectra), astrocytomas grade II (Class 2, 22 spectra), glioblastomas (87 spectra) and metastases (39 spectra) [6]. However, the last two types of tumor are very difficult to distinguish. Our experience on this dataset is that, the trained models did not perform as well as a majority classifier, which assigns the majority class in the training set to all the test cases. Therefore, we merged the two tumor types - glioblastomas and metastases - into one class of aggressive tumors (Class 3), and only dealt with the three-class classification problem. For details of the data acquisition and preprocessing procedure for this dataset, the readers are referred to [6].

Since the data are unbalanced, the model using the default priors will lead to a relatively low sensitivity for astrocytomas grade II. Thus, we decided to use equal priors for all binary classifiers, which resulted in a "satisfactory" sensitivity and specificity for all three classes. Table III reports the average test AUC for each pairwise binary classification, and Table IV presents both the training and test accuracy of the brain tumor classification problems using equal class priors. Again, in these tables, the highest AUC of each binary classification for each type of model (in row) is indicated in bold.

For the pairwise binary classification, similar observations can be found as with the microarray data. For the 3-class brain tumor diagnosis, as reported in Table IV, LinSBL+Bag got a consistently higher accuracy than the rest of variable selection methods. Also, the paired t-tests on accuracy indicate that LinSBL+Bag performed significantly better than LinSBL with a p-value $< 10^{-4}$ for each model type.

D. Biological relevance of the selected variables

We examined the most frequently selected variables from the LinSBL+Bag method and their biological relevance for each dataset. The heatmap in Fig. 1 and 2 shows the occurrence of such highly selected genes in each randomized cross-validation for the leukemia data and the colon cancer data, respectively.

It is noteworthy that the genes that were selected by the LinSBL+Bag method are mostly biologically interpretable. In the leukemia cancer classification, the three top ranked genes identified by our algorithm are all among the informative genes according to [9]. The highest ranked gene is zyxin (gene 4847 according to its order in the original dataset), which encodes an LIM domain protein localized at focal contacts in adherent erythroleukemia cells. CD33 (gene 1834) is the differentiation antigen encoding cell surface protein, for which monoclonal antibodies have been demonstrated to be useful in distinguishing lymphoid from myeloid lineage cells.

In the colon cancer data, the most important gene which is identified by our method corresponds to mRNA for uroguanylin precursor (gene 377). Guanylin and uroguanylin have been recently found to be linked to colon cancer, and treatment with uroguanylin was found to have possible therapeutic significance [15][20]. The gene with the second highest rank (Gene 1772) is a collagen alpha2 (XI) chain which is involved in cell adhesion, and collagen degrading activity is part of the metastatic process for colon carcinoma cells [10][14].

For the class prediction of brain tumors, we examined the corresponding metabolites of which the magnitude values

TABLE III

TEST AUC (%) FOR PAIRWISE BINARY CLASSIFICATION OF BRAIN TUMORS.

VS method	All			Fisher+CV			RFE+CV			LinSBL			LinSBL+Bag		
Class pair	1vs.2	1vs.3	2vs.3	1vs.2	1vs.3	2vs.3	1vs.2	1vs.3	2vs.3	1vs.2	1vs.3	2vs.3	1vs.2	1vs.3	2vs.3
Classifier $\overline{N}_{\rm v}$	138	138	138	39.17	114.57	9.03	6.10	23.10	14.07	4.30	9.50	6.20	31.17	59.67	55.40
SVM	**99.82**	97.52	92.14	98.75	96.17	95.06	98.18	96.56	92.23	96.99	96.51	90.78	98.17	**97.76**	**96.01**
	±0.08	±0.26	±0.77	±0.41	±0.89	±0.69	±0.56	±0.37	±0.95	±0.62	±0.37	±1.07	±0.88	±0.28	±0.44
BayLSSVM	99.62	97.33	95.15	98.67	96.85	95.43	97.44	97.04	94.23	96.94	96.88	93.38	**99.65**	**97.95**	**96.44**
	±0.16	±0.27	±0.56	±0.39	±0.39	±0.63	±0.57	±0.30	±0.70	±0.57	±0.34	±0.82	±0.15	±0.26	±0.40
RVM	98.47	97.55	**96.87**	98.60	96.82	95.80	97.52	97.18	95.57	96.99	97.02	94.52	**99.70**	**97.94**	96.19
	±0.37	±0.27	±0.37	±0.38	±0.47	±0.65	±0.74	±0.32	±0.67	±0.72	±0.35	±0.70	±0.13	±0.26	±0.42

TABLE IV

TRAINING AND TEST ACCURACY FOR BRAIN TUMOR THREE-CLASS CLASSIFICATION.

VS method	All		Fisher+CV		RFE+CV		LinSBL		LinSBL+Bag	
$\overline{N}_{\rm v}$	138		115.97		37.37		17.9		98.73	
Classifier	Train(%)	Test(%)	Train(%)	Test(%)	Train(%)	Test(%)	Train(%)	Test(%)	Train(%)	Test(%)
SVM	100.00	85.25	95.72	85.54	100.00	85.05	99.95	83.92	96.40	**86.91**
	±0.00	±0.69	±1.07	±1.10	±0.00	±0.70	±0.05	±0.43	±0.30	±0.71
BayLSSVM	99.25	86.37	94.57	86.47	98.54	85.15	96.20	86.37	96.84	**89.51**
	±0.12	±0.75	±0.52	±0.66	±0.18	±0.66	±0.25	±0.79	±0.20	±0.55
RVM	89.17	87.72	89.49	87.11	94.55	87.11	94.60	86.72	94.67	**89.95**
	±0.31	±0.76	±0.34	±0.61	±0.43	±0.70	±0.38	±0.76	±0.19	±0.56
C4.5	98.10	78.75	95.77	80.15	98.00	79.90	97.49	79.90	99.05	**84.02**
	±0.17	±0.73	±0.46	±0.70	±0.16	±0.72	±0.23	±0.80	±0.13	±0.70

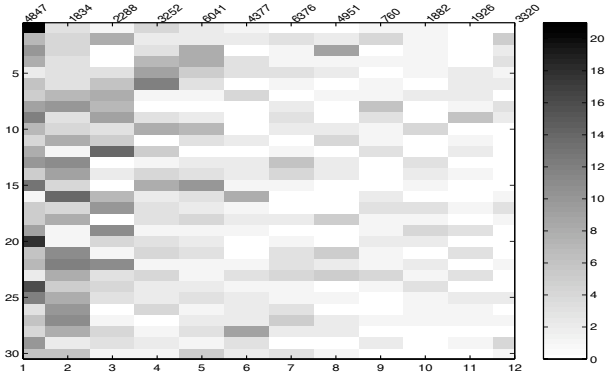

Fig. 1. Genes selected by LinSBL+Bag from the 30 realizations of the training sets for leukemia cancer microarray data. The x-axis labels in the bottom the rank of the gene, and on the top the index of the gene in the original microarray data matrix. The y-axis refers to the run number in the 30 randomized cross-validations. Only the genes that were selected more than 30 times in all the 30×30=900 linear SBL models are listed in the plot. The gray level in each cell corresponds to the number of occurrences that a gene was selected in bootstrapping for one realization of the training set.

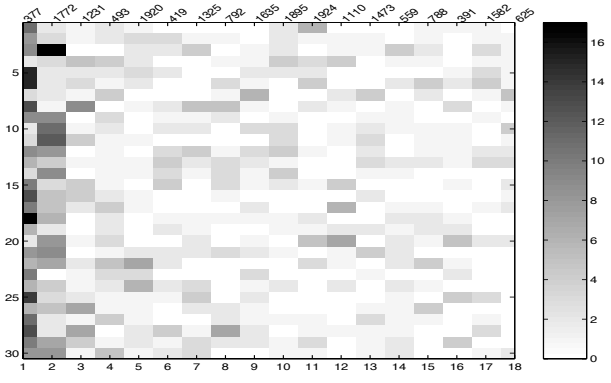

Fig. 2. Genes selected by LinSBL+Bag from the 30 realizations of the training set for colon cancer microarray data. See Fig. 1 for details.

appeared to be significant in the pairwise binary classification problems. Fig. 3 depicts the mean spectra of the three classes, as well as the metabolites associated with their resonance frequencies [8]. The two horizontal bars, below each mean spectrum, represent the selection rate of each variable in pairwise discrimination with another tumor class. The selection rate for variable m was computed by dividing the total number of selection occurrences by $30 \times 30 = 900$. We can take the selection rate as a means of importance measure for the variables. By examination of the figure and incorporating the domain knowledge, we were able to figure out the metabolites that are important or useful for the classification.

For example, in contrast to the other two classes, the astrocytomas grade II have a relatively high level[6] in the frequency regions of both total creatine (Cr) and myo-inositol (mI)/ glycine (Gly). These variables were also selected most frequently in all three pairwise binary classification problems, particularly for differentiating Class 1 from 2. Indeed, in these regions the selection rate has the darkest color and reaches a value close to 0.5. To discriminate meningiomas from the aggressive class, more frequency regions are used: not only Cr and mI/Gly, but also Glutamate (Glu), Glutamine (Gln), Lipids, N-acetyl containing macromolecules (NAC). Interestingly, Cr does play a role in the maintenance of energy metabolism. While NAC resonances at the usual NAA (N-acetyl aspartate) chemical shift may appear in the solid or cystic areas of brain tumors.

However one must be cautious when interpreting the selected variables for such MR spectra: the resonances at the same position may originate in different compounds depending on the tumor type. For example, the 2.03 ppm peak originates mostly from Lipids in Class 3 tumors, while it is safe to label it NAC for Class 1 and 2. It may have NAA contribution, but other N-acetyl compounds are contributing varying amounts [5]. The whole region at 2-2.6 ppm may have variable contribution from macromolecules (mostly proteins).

[6]The "level" here refers to relative intensity in the spectra as they are scaled to unit norm.

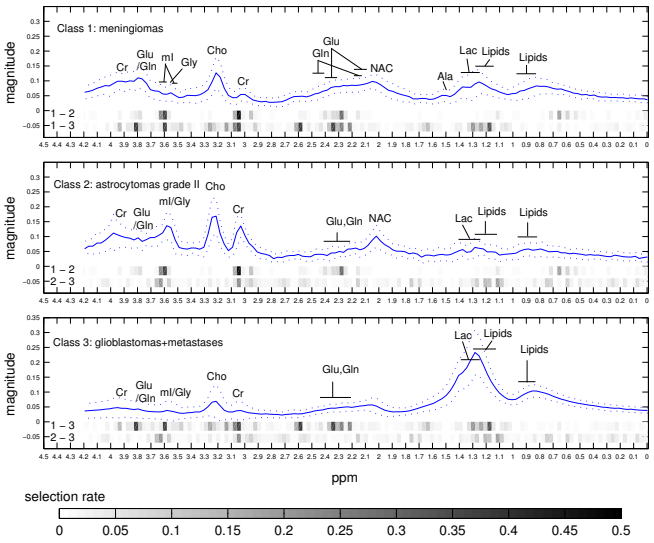

Fig. 3. Mean spectrum of the brain tumors and the selection rate of the variables using LinSBL+Bag from the 30 runs of cross-validation for the pairwise binary classification. The dotted line is the mean $\pm$ SD spectrum.

VII. Discussion

From both the binary and multiclass examples, we can clearly see how bagging improves the performance of the single model using only one subset of variables. There is a significant gap in predictive performance between the methods of LinSBL and LinSBL+Bag in our experiments. This gap might result from: on the one hand the large uncertainty because of the small sample size of the training data, and on the other hand the sensitivity of SBL itself.

The results also show that models with only a small portion of variables can perform as well as, and sometimes even better than the models with the complete set of variables. More importantly selected variables could help in interpreting and understanding the underlying mechanism of the diseases.

As to the modelling techniques, the kernel-based models performed consistently better than the decision-tree models. And the Bayesian probabilistic models performed somehow better than the standard SVMs. This might be partially due to the fact that the hyperparameters for SVMs were not optimized. Our main focus was on the models with probabilistic output which is important in biomedical diagnosis, without the burden of cross-validation for hyperparameter tuning.

It is also worth mentioning that the linear SBL model is itself a probabilistic model with an automatic variable selection mechanism. It achieved a similar performance as the probabilistic kernel models in all these experiments. For example, in the brain tumor diagnosis problem, a single LinSBL model and a bagged LinSBL model yielded a mean test accuracy of $86.03\pm0.63\%$ and $89.46\pm0.52\%$, respectively.

To get an idea of the computational efficiency of our VS methods, we computed the average CPU time in a CV trial consumed by LinSBL+Bag on 30 bootstrap training samples. The simulations were conducted on cluster machines with Pentium processors (1GHz). Around 7 minutes and 3 minutes were needed for the leukemia data and the colon data, respectively. For prediction of brain tumors, in total 24 minutes were used for all 3 pairs of binary classification problems.

One limitation of the bagging strategy is that there is no single model to be returned. To deal with this problem, one can adapt a similar approach as described in [22], in which the linear discriminant analysis (LDA) was bootstrapped, to generate an "averaged" classifier using the weighted average of the B sets of model parameters. To do so, a transformation from our linear kernel-based models to variable based models could be done and the selection frequency for the variables should also be taken into account. One direction for future investigation could be to establish a mechanism for integrating multiple models into a single structure model, which would become easy to explain clinically.

VIII. Conclusions

The most significant problem for classification addressed here lies in the use of datasets with a small sample size and a huge dimensionality. The populations are usually underrepresented in this situation, which might result in a serious bias towards the training set, i.e. with a high training performance for a single model after variable selection and possibly a much lower generalization performance on the unseen data. This motivates the use of a bagging strategy in order to improve the reliability and lower the uncertainty in both variable selection and modelling.

Experimental results confirm the advantages of the bagging strategy. Indeed, bagging can enhance the reliability of variable selection and model prediction, thereby increasing the generalization performance of the models. Unlike the popular variable ranking methods such as RFE and the Fisher's criterion methods, our proposed method requires no additional step in order to decide on the number of variables to be used in the models. The variables are selected within a Bayesian framework, and the procedure is shown to be computationally efficient if the sample size is small. The number of occurrences of a variable being selected can serve as an importance measure for the variable. Our results imply that the linear sparse Bayesian learning plus bagging deserves to play an important role in variable selection for biomedical classification tasks.

Acknowledgments

The authors would like to thank the anonymous reviewers for the detailed and valuable comments. Use of the brain tumor data provided by the EU funded INTERPRET project (IST-1999-10310, http://carbon.uab.es/INTERPRET) is gratefully acknowledged.

References

[1] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 6745-6750, 1999.

[2] C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[3] C.M. Bishop, and M.E. Tipping, Bayesian regression and classification. In J.A.K. Suykens *et al.* (Eds.), *Advances in Learning Theory: Methods, Models and Applications*, vol. 190, NATO Science Series III: Computer and Systems Sciences, IOS Press, pp. 267-288, 2003.

[4] L. Breiman, Bagging predictors, *Machine Learning*, 24:123-140,1996.

[5] A.P. Candiota, C. Majós, A. Bassols, M.E. Cabañas, J.J. Acebes, M.R. Quintero, C. ArúS, Assignment of the 2.03 ppm resonance in in vivo ^{1}H MRS of human brain tumour cystic fluid: contribution of macromolecules, *MAGMA*, vol. 17, pp. 36-46, 2004.

[6] A. Devos, L. Lukas, J.A.K. Suykens, L. Vanhamme, A.R. Tate, F.A. Howe, C. Majós, A. Moreno-Torres, M. Van der Graaf, C. Arús, S. Van Huffel, Classification of brain tumours using short echo time ^{1}H MR spectra, *Journal of Magnetic Resonance*, 173(2):218-228, 2004.
[7] L.M. Fu, E.S. Youn, Improving reliability of gene selection from microarray functional genomics data, *IEEE Transations on Information Technology in Biomedicine*, vol. 7, no. 3, pp. 191-196, 2003.
[8] V. Govindaraju, K. Young, and A. A. Maudsley, Proton NMR chemical shifts and coupling constants for brain metabolites, *NMR in Biomedicine*, vol. 13, pp. 129-153, 2003.
[9] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M. A. Caliguiri, C. D. Bloomeld, and E.S. Lander. Molecular classication of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, vol. 286, pp. 531-537, 1999.
[10] I. Guyon , J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, *Machine learning*, vol. 46, pp. 389-422, 2002.
[11] J.A. Hanley, B. McNeil, The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve, *Radiology*, vol. 143, pp. 29-36, 1982.
[12] T. Hastie, R. Tibshirani, Classification by pairwise coupling. In M.I. Jordan, M.J. Kearns, and S.A. Solla, editors, *Advances in Neural Information Processing Systems*, vol. 10, MIT Press, 1998.
[13] T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning - data mining, inference, and prediction*, Springer, New York, 2001.
[14] G. Karakiulakis, C. Papanikolaou, S.M. Jankovic, A. Aletras, E. Papakonstantinou, E. Vretou, V. Mirtsou-Fidani, Increased type IV collagen-degrading activity in metastases originaing from priorary tumors of human colon *Invasion and metastasis*, 17(3):158-168, 1997.
[15] Y. Li, C. Campbell, M. Tipping, Bayesian automatic relevance determination algorithms for classifying gene expression data. *Bioinformatics*, vol. 18, pp 1332-1339, 2002.
[16] S.K. Mukherji (ed.), *Clinical Applications of Magnetic Resonance Spectroscopy*. Wiley-Liss, 1998.
[17] S.J. Nelson, Multivoxel magnetic resonance spectroscopy of brain tumors, *Molecular Cancer Therapeutics*, vol. 2, pp. 497-507, 2003.
[18] A.E. Nikulin, B. Dolenko, T. Bezabeh, R.L. Somorjai, Near-optimal region selection for feature space reduction: novel preprocessing methods for classifying MR spectra, *NMR Biomed*, vol. 11, pp. 209-216, 1998.
[19] J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, Inc., 1993.
[20] K. Shailubhai, H.H. Yu, K. Karunanandaa, J.Y. Wang, S.L. Eber, Y. Wang, N.S. Joo, H.D. Kim, B.W. Miedema and S.Z. Abbas *et al.*, Uroguanylin treatment suppress polyp formation in the Apc$^{Min/+}$ mouse and induces apoptosis in human colon adenocarcinoma cells via cyclic GMP. *Cancer Res.*, vol. 60, pp. 5151-5163, 2000.
[21] R. Simon, M. Radmacher, K. Dobbin, L. McShane, Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst*, vol. 95, pp. 14-18, 2003.
[22] R.L. Somorjai, B. Dolenko, A. Nikulina, P. Nickersonb, D. Rushb, A. Shawa, M. Glogowskia, J. Rendella, R. Deslauriers, Distinguishing normal from rejecting renal allografts: application of a threestage classification strategy, *Vibr. Spectrosc*, vol. 28, pp. 97-102, 2002.
[23] J.A.K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, *Neural Processing Letters*, vol. 9, no. 3, pp. 293-300, 1999.
[24] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, *Least Squares Support Vector Machines*. Singapore: World Scientific, 2002.
[25] M.E. Tipping, Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, vol. 1, pp. 211-244, 2001.
[26] M.E. Tipping and A. Faul, Fast marginal likelihood maximisation for sparse Bayesian models. *Proc. Artificial Intelligence and Statistics*, 2003.
[27] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.
[28] T. Van Gestel, J.A.K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, J. Vandewalle, A Bayesian framework for Least Squares Support Vector Machine classifiers, Gaussian processes and kernel Fisher discriminant analysis," *Neural Computation*, vol. 14, pp. 1115-1148, 2002.
[29] J. Weston, A. Elisseeff, M. Tipping, B. Schölkopf, Use of the zero norm with linear models and kernel methods. *Journal of Machine Learning Research*, vol. 3, pp. 1439-1461, 2002.
[30] M. Xiong, X. Fang, J. Zhao, Biomarker identification by feature wrappers. *Genome Res.*, vol. 11, pp. 1878-1887, 2001.

Chuan Lu received the B.E. degree in management information system from the Tianjin University, China in 1995, the master's degree in artificial intelligence from K.U.Leuven, Belgium in 2000 and the PhD degree from the Department of Electrical Engineering, K.U.Leuven in 2005. She is currently a research associate at Computational Biology Group, the Department of Computer Science, in the University of Wales, Aberystwyth, UK. Her research interests focus on statistics and machine learning applied to biology and medicine.

Andy Devos received the master's degree in mathematics (Katholieke Universiteit Leuven 1999) and in artificial intelligence (Katholieke Universiteit Leuven 2000), and the PhD degree in engineering (the Department of Electrical Engineering, K.U.Leuven 2005). The topic of his PhD was MRS data analysis with applications to brain tumour recognition.

Johan Suykens received the degree in electromechanical engineering and the Ph.D. degree in applied sciences from the K. U. Leuven, in 1989 and 1995, respectively. In 1996 he was a Visiting Postdoctoral Researcher at the University of California, Berkeley. He has been a Postdoctoral Researcher with the Fund for Scientific Research FWO Flanders and is currently a Professor with K.U.Leuven. His research interests are mainly in the areas of the theory and application of neural networks and nonlinear systems, in which he (co)authored over 300 papers, three books and edited two books.

Dr. Suykens is a Senior IEEE member and has served as associate editor for the IEEE Transactions on Circuits and Systems- Part I (1997-1999) and Part II (since 2004) and since 1998 he is serving as associate editor for the IEEE Transactions on Neural Networks. He received an IEEE Signal Processing Society 1999 Best Paper (Senior) Award and several Best Paper Awards at International Conferences. He is also a recipient of the International Neural Networks Society INNS 2000 Young Investigator Award for significant contributions in the field of neural networks.

Carles Arús received the BSc degree in biology from the Universitat Autónoma de Barcelona (UAB), Spain in 1976, and the PhD degree in chemistry UAB in 1981 on the subject of the sub-site structure of bovine pancreatic RNase A (enzyme kinetics, NMR spectroscopy). He received the Best thesis award in the Faculty of Sciences of UAB in 1982. He was a postdoctoral researcher in the USA (1982-1985) at Univ. Illinois, Chicago, IL and Purdue Univ., IN. He is currently a "Catedrático" (full Professor) in the Department of Biochemistry and Molecular Biology of the UAB.

The research group of Dr. Arús has carried out work on the application of NMR spectroscopy of tumours for diagnostic purposes and has also contributed to the investigation of human muscle bioenergetics by ^{31}P MRS. His present research interests are in the field of tumour spectroscopy target the use of ^{1}H MRS of human brain tumours, biopsies and cell models for diagnosis, prognosis and therapy planning. He has published 56 PubMed accessible articles since 1977.

Sabine Van Huffel received the MD in computer science engineering in June 1981, the MD in Biomedical engineering in July 1985 and the Ph.D in electrical engineering in June 1987, all from K.U.Leuven, Belgium. She is full professor at the department of Electrical Engineering from the Katholieke Universiteit Leuven, Leuven, Belgium. Her research interests are in signal processing, numerical linear algebra, errors-in-variables regression, system identification, pattern recognition, (non)linear modelling,software, statistics, applied to biomedicine. In these areas, she has authored and co-authored 2 books, 2 edited books, more than 150 papers in International Journals, 4 book chapters, and more than 140 conference contributions.